Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 17 (2013) 844 – 851

Procedia Computer Science

Information Technology and Quantitative Management (ITQM2013)

A multiple watermarking algorithm for texts mixed Chinese and English

Xu Rui[a,b,*], Chen XiaoJun[b], Shi Jinqiao[b]

[a]*Beihang University, Beijing 100191, China*
[b]*Institute of Information Enginerring, Chinese Academy of Sciences, Beijing 100190, China*

Abstract

Digital watermarking is a common digital product copyright protection mechanism, and used widely in information security research area. But there are still some deficiencies in current digital text watermarking algorithm, such as low robustness, insufficient security, low watermarking capacity and inability of application to texts mixed Chinese and English. The paper proposes a multiple watermarking algorithm for texts mixed Chinese and English based on character encoding and attributes. This algorithm can provide a larger watermarking capacity, and meanwhile keep the watermarking information invisible. And then it also provides strong security and auto error correction ability to guarantee that the watermarking is not destroyed or tampered by other malicious attackers. Our experiment and related analysis demonstrate all these features.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords:Text digital watermarking; Texts mixed Chinese and English; Robustness; Security

1. Introduction

Digital watermarking is a digital product copyright protection technology which is proposed in the 1990 s, in the information security research field. It is embedded in the multimedia data, digital, serial number, text, image, logo and other types of information, and plays a role of copyright protection, signs product, secret communications, confirming data belonging, identifying data authenticity and so on [1].

Nowadays the main text watermarking algorithms can be classified in 4 ways: the text watermarking based on format, the text watermarking based on content, the text watermarking based on unimportant represented, and the text watermarking based on natural language [2].

The text watermarking based on format is the most widely used algorithm. Row shift, character shift from the initial, later the feature coding to progress which alter the font size, color and other methods [3, 4], and the research of text watermarking based on format is very active.

The text watermarking based on content is replaced by character encoding in order to achieve the purpose of watermarking embedding, which has better robustness and security. Literature [5] uses similar Greek letters in shape to replace the English letters, but this method is only suitable for the English. And the watermarking space is small and cannot embed more watermarking information. Literature [6] the text watermarking based on the Chinese character expression is the best kind of watermarking algorithm for the Chinese text. But the algorithm is only

* Corresponding author. Tel.: +86-10-82546735.
E-mail address: xurui@nelmail.iie.ac.cn.

Selection and peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management
doi:10.1016/j.procs.2013.05.108

applicable to the Chinese text.

The text watermarking based on unimportant represented is similar to the LSB (Least Significant Bit) of the image watermarking, which use the unimportant punctuation position or spaces to embed watermarking. Such watermarking algorithms are unstable and possible to lost watermarking information in the transmission process. Moreover robustness and security are not as good as the later text watermarking based on format.

The text watermarking based on natural language was presented earliest in the Purdue University by Mikhail.J.Atallah and VictorRaskin et al [7] in the United States in 2002. The watermarking information is embedded by changing the sentence structure, synonym substitution and other methods. Naturally language digital watermarking changed the text content, but it hardly changed the meaning of the text and format. After adding watermarking, it is almost impossible to detect watermarking information in the file. Also the watermarking is not easy to be destroyed. But for standard file, this method may change the semanteme, because its format requirement is strict. Therefore this method does not apply to format demanding file. And because the computer natural language processing is not enough mature, this is the bottleneck of the text watermarking based on natural language technology.

The existing watermarking algorithms have many problems, such as:

- Low robustness: we can operate directly through WORD menu to alter the watermark information;
- Insufficient security: if you know watermarking algorithm, then you can extract the watermarking information;
- Low watermarking capacity;
- Cannot apply to texts mixed Chinese and English and so on.

In this paper, a new watermarking algorithm has been proposed which is based on format text watermarking, is the multiple watermarking algorithms for texts mixed Chinese and English, it based on character encoding and attributes. This algorithm improves the watermarking security, robustness and capacity better than the existing watermarking algorithm. Furthermore the new watermarking algorithm can be put into practical use already.

The remainder of the paper is organized as follows: Section 2 describes watermarking algorithm in detail. In section 3 we describe and analysis experimental results. Then we conclude the paper and provide future work in section 4.

2. Watermarking algorithm

2.1. Overview

We combine watermarking information with key, after code conversion and stratified. We can use MD5 to process original watermarking, and then embed watermarking information in order to generate the embedded watermarking file. In Figure 1, we give out the model of watermarking embedding.

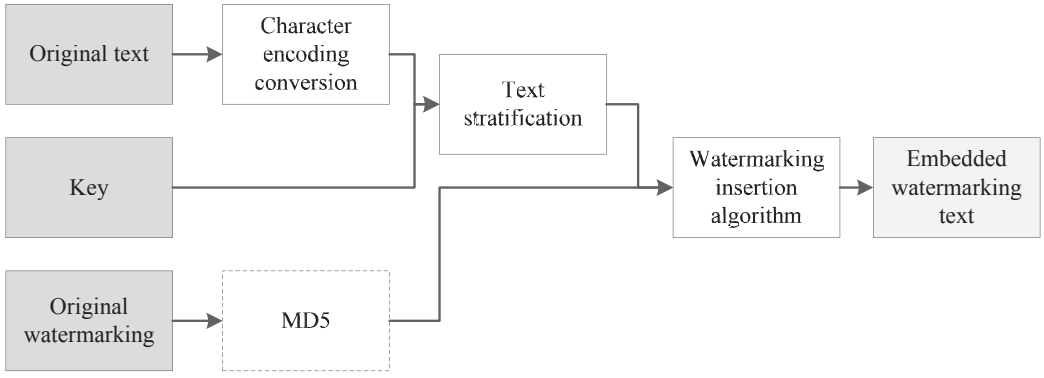

Fig. 1. Watermarking embedding model.

2.2. Watermarking insertion algorithm

A. Watermarking generation

In order to enhance robustness of the watermarking, we proposed multiple watermarking embedding methods based on the characteristics of the WORD document. However, the multiple watermarking embedding method will increases demands for the text watermarking capacity. When the amount of the embedded watermarking information is slightly larger, the method has higher requirement to the text length. Based on the consideration of lowering text length limit and the amount of original watermarking information as little as possible, we choose a safe and reliable function of HASH (MD5 algorithm) to obtain the original watermarking information (128 bits) as the final loaded watermarking data. On the one hand, the original watermarking play encrypted effect. On the other hand, the length of the original watermarking information may not be restricted. Accordingly we use the loaded watermarking data and original watermarking information as table records to establish database table in order to get the original watermarking information after extracting the embedded watermarking [8].

B. Character encoding conversion

We choose every character as one section. Every Chinese character encoding is 2 byte, but 1 char has only 1 byte. So we should switch character encoding to wide character encoding. In our experiment, the wchar_t character data type is equivalent to Unicode encoding, and we use this character encoding to operate.

C. Multiple watermarking embedding method

We use two invisible attributes of the WORD document to embed the watermarking information [8].

a) Based on the NoProofing attributes of characteristics, default value of the Selection object WORD (such as character) remains unchanged embedded watermarking data and can only be modified through programming, so we use NoProofing attribute of character to embed watermarking data. Its value is TRUE if the spelling and grammar checker ignores the specified text. It returns wdUndefined if the NoProofing property is set to TRUE for only some of the specified text. Read/write Long.

b) Using LanguageIDOther attribute of characters, we set this attribute for the use of a number of less language enumeration values. Microsoft recommended way to return or set the language of Latin text in a document created in a right-to-left language version of Microsoft Word. This property has a total of 64 enumeration values. After sifted, select which three values (wdBasque 1069, wdVenda 1075, wdEstonian 1061) as the modified value. In that way each character can be embedded two bits watermarking information so as to improve the watermarking capacity.

The two attributes of characters can't be found, added and modified except for programming. Meanwhile WORD menu operation can't clear watermarking characteristics, neither. Thus it is strongly concealed and resistant to attack. We can modify the character of these two properties simultaneously, but in order to improve the robustness and security, this algorithm only change one of the properties of each character after text has been stratified.

D. Text stratification

Take the last two bits of key characters encoding as the actual key. Meanwhile we take the last two bits of text character Unicode encoding, and get the value of XOR between them, if

Results for 00 and 10, is classified as the first layer, modify NoProofing bit.

Results for 01 and 11, is classified as the second layer, modify LanguageIDOther bit.

E. Watermarking insertion algorithm

We know that every character of the second layer can be embedded two bits. Repeat N times in each layer in order to improve the robustness, and add a separator between each group. We choose a Unicode operational

character as a separator, which is RLO (Right-to-Left Override), and its binary sequence is 10000000101110. N is equal to the floor value of the total number of characters divided by 128. Then we embed watermarking according to the following steps.

i. Import the original watermarking content, key and pending text;
ii. Use the MD5 algorithm to obtained a summary of the data of the original watermarking as the finally loaded watermarking data;
iii. Divide all the characters of the full text into two layers;
iv. According to the order from front to back on the first layer of words embedded N groups watermarking.
v. According to the order from back to front on the second layer of words embedded 2*N groups watermarking.

And the watermarking embed rule as shown below with pseudocode segment.

```
if (whether current character is in the first layer)
  if (current watermarking bit == 1)
        NoProofing = TRUE;
end
else
  switch (current watermarking bits)
        case 01: LanguageIDOther = wdBasque; break;
        case 10: LanguageIDOther = wdVenda; break;
        case 11: LanguageIDOther = wdEstonian; break;
        default: break;
  end
end
```

2.3. Watermarking detection and extraction algorithm

a) Watermarking detection

The default values of NoProoing and LanguageIDOther are FALSE and 1033 (wdEnglishUS). By detecting the input text character, if one of the values of these two properties is not the default value, then the document is an embedded watermarking file.

b) Watermarking extraction

We can do the following steps to extract the watermarking information.

i. Divide text character in process into two layers by the rules.
ii. According to the rules of each layer to extract the embedded watermark, get 3*N groups of watermarking.
iii. When the 3*N groups of watermarking are consistent and match a record of the database, it indicating that all the watermarking are completely normal. Then file has not any signs of damage, namely the integrity of the text is not destroyed, and watermarking extraction is completely normal, so we reach the end of the program. Otherwise, turn to the watermarking error correction algorithm.

2.4. Watermarking error correction algorithm

If the embedded watermarking file has been attacked, such as deleting or adding character, we can do the following steps try to get the correct watermarking information.

i. The 3*N group of watermarking are inconsistent, but at least one group of watermarking matching database record, suggesting that the copyright information and prompting the damage of files integrity based on the type of error found. Otherwise, turn ii.
ii. The all 3*N group of watermarking does not match the database records, we use 3*N groups of watermarking to complement each other, check and error correction, to restore one group of complete watermarking, prompting extracted copyright and the damage of file integrity information. Otherwise, turn iii.
iii. This file is damaged, and we cannot recover the watermarking.

2.5. Summary

The watermarking information hidden in the carrier text cannot be easily found. This algorithm uses a key to realize the secondary hidden when hiding watermarking information. In the text layered process, hiding the location and the corresponding attributes of watermarking embedding characters may not be changed. If the attacker does not understand this algorithm, just extract the corresponding attributes of characters, he will not know what position the watermarking embedding and cannot get the watermarking information except that he crack the key . Then this algorithm uses the MD5 value of original watermarking information to further improve the security. Even if the attacker can obtain the embedded information, he still doesn't know the actual watermarking information. Finally, this algorithm can embed more than one bit into one character, and it has error correction process. Therefore the multiple text watermarking algorithm, has high security, high capacity and high robust.

3. Experimental results and analysis

Our experiment environment is windows XP, Visual C++ 6.0 and Microsoft WORD 2010.

3.1. Watermarking embedding experiment

Experimented with a file which been mixed Chinese and English words. The following diagrams show the contrast before and after the experiment.

近日，有消息报道，微软指控凯里森建筑为其华盛顿中心提供了劣质的室内系统设计。微软表示，凯里森的失职将可能导致该大楼内部的漏水问题，进而危及到放置在里面的近千台服务器的运行。As a high-performance hard drive alternative, Intel Solid-State Drives boost your PC to the next level in storage performance and reliability. Quality you can trust. Sounds great, now determine which Intel SSD is right for your computing needs.↵

Fig. 2. Original file.

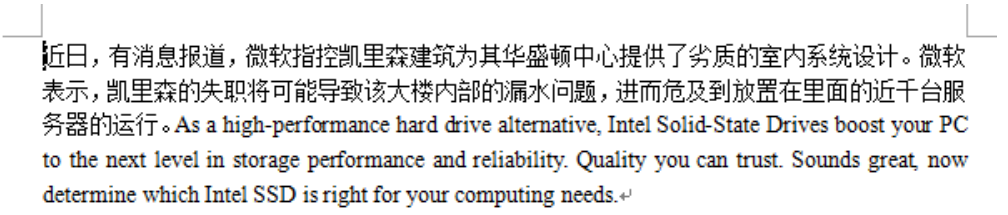

Fig. 3. Embedded watermarking file.

It is obvious that this algorithm has a well invisible of watermarking and no influence on the visual effect of text.

3.2. Watermarking capacity analysis

According to our method of text stratification, the watermarking capacity of the first layer should be about 50% theoretically. It means we can embed 50 bits into every one hundred characters. So we conduct an experiment to prove this.

Table 1. The watermarking capacity of the first layer

The total number of characters	The number of first layer and second layer characters	The number of first layer characters/ The total number of characters	The number of second layer characters/ The total number of characters	The total number of watermarking bit
87	37(50)	42.52%	57.48%	157.48%
158	106(52)	67.09%	32.91%	132.91%
230	123(107)	53.48%	46.52%	146.52%
156	76(80)	49.06%	50.94%	150.94%
180	112(68)	62.22%	37.78%	137.78%
258	158(100)	61.24%	38.76%	138.76%
183	73(110)	39.89%	60.11%	160.11%
167	98(69)	58.68%	41.32%	141.32%
155	74(81)	47.74%	52.26%	152.26%
222	99(123)	44.59%	55.41%	155.41%
Averages：		52.65%	47.35%	147.35%

The table above shows, the watermarking capacity of the first layer which modified the value of NoProofing bit is 52.65%, which is close to the theoretical data. While the second layer which modified the value of LanguageIDOther bit, for each character can be embedded into two information bits, the watermarking theoretical capacity is about 100%, and its actual value is 47.35*2%. The total watermarking theoretical capacity of this algorithm is about 150%, and its actual value is 147.35%.

In Table 2 is shown the existing large text watermarking algorithm capacity. It indicates the multiple texts watermarking algorithm for texts mixed Chinese and English have the highest watermarking capacity.

Table 2. The existing large text watermarking algorithm capacity

Text watermarking algorithm	Watermarking capacity
The multiple texts watermarking algorithm based on the Unicode encoding	150
The text watermarking algorithm based on tone and character feature	114
The text watermarking algorithm based on character space	50
The text watermarking algorithm based on character feature	50
The text watermarking algorithm based on character structure	Less than 50
The text watermarking algorithm based on tone	25

3.3. Watermarking security analysis

Using embedded binary sequence 0100111111100001 into file as an example, and the key is 0110000001101111. We do not use MD5 as the watermarking information, because it is easier to explain with a short data. The actual key is the last two bits of original key, namely 11. After embedded watermarking into file, the NoProofing bits sequence as shown below.

```
0100111111100001001000000010111001001111111000010010000000101110010011111110
0001001000000010111001001111111000010010000000101110010011111110
```

Fig. 3. The NoProofing bits sequence of the embedded watermarking file

If a error key is inputted, the extracted information will sequence as shown below. The positions of embedded watermarking are wrong, then the extracted information is completely wrong, so we cannot get the real watermarking information. So this algorithm has a higher security.

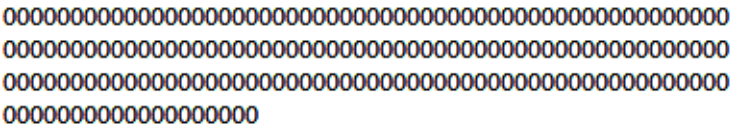

Fig. 4. The extracted wrong NoProofing bits sequence.

3.4. Watermarking error correction analysis

We delete 40 characters in the embedded watermarking file (the total number of file characters is 124), and then extract the watermarking information. The NoProofing bits sequence as shown below. The original watermarking information is 0100111111100001, and the separator is 10000000101110.The sequence in box is the watermarking, and the sequence with underline is the separator. Then we can get 4 groups of watermarking and 2 groups of separator. The first and second watermarking is useless, because they do not have separator before them. But we can get the complete watermarking information through the useful watermarking sequence complement with each other. Finally the original watermarking information been restored.

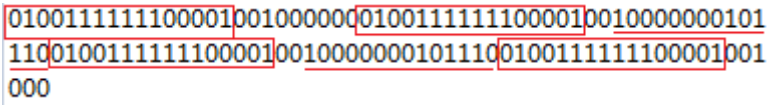

Fig. 5. The schematic diagram of watermarking error correction.

4. Conclusions

In this paper, this proposed algorithm has the following features. It has strong concealed, which is difficultly detected by users, and the possibility to be destroyed is low. Watermarking information use MD5 for encrypt and improve the security. Use a key when watermarking embedding to further improve the watermarking security. Even if someone knows this algorithm, he still need crack key to obtain watermarking information. When watermarking embedding the second layer, each character can be embedded in two information bits that improves the watermarking capacity. If increase the watermarking embedding position, the watermarking capacity can reach 200%, the special circumstances shows the large watermarking capacity of this algorithm. When text is under various attacks, this algorithm has stronger watermarking extraction and error correction ability. It also can be practicable for texts mixed Chinese and English.

For future work, we plan to do the following things. First of all, although we use the character Unicode encoding as the key, the algorithm can still be cracked extremely easily because of only two binary bits in use, and the repeatability is also very high. Hence we consider further improving the key feature to enhance security. Second, the error correction algorithm needs to be further enhanced. Third, the LanguageIDOther property has 64 enumeration values. By further screening we choose more kinds of useful values, thereby enhancing the watermarking capacity greatly. Last but not least, this algorithm is only available to Word format now. The next step is supposed to let the algorithm support more formats, such as Excel, WPS, PDF, and so forth.

Acknowledgements

This work is supported by National High Technology Research and Development Program of China, 863 Program (Grant No.2011AA010701), National Key Technology R&D Program (Grant No.2012BAH37B04), and Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA06030200).

References

1. Cox I J, Linnartz J P M G. Public Watermarks and Resistance to Tampering[C]. Proc. of International Conference on Image Processing, Santa Barbara, California, 1997-10: 26-29.

2. ZHAO Jie (2007). A text watermarking algorithm for texts mixed Chinese and English. Journal of Communication and Computer 4(9): 9-13.

3. Brassil J T, Low S, Maxemchuk N F. Copyright Protection for the Electronic Distribution of Text Documents[J]. Proceedings of the IEEE, 1999, 87(7).

4. Brassil J, Low S, Maxemchuk N F, et al. Electronic Marking and Identification Techniques to Discourage Document Copying[J]. IEEE Journal on Sel. Areas in Commun, 1995, 13(8): 1495-1504.

5. Xiao, X. and X. Sun (2005). Design and Implementation of Content-based English Text Watermarking Algorithm. Computer Engineering 31(25): 29-31.

6. SUN X.M., LUO G., HUANG H.J. Component-based digital watermarking of Chinese texts. Proc. of the 3rd International Conference on Information Security. ACM International Conference Proceeding Series, 2004, 85.

7. Atallah M J, Raskin V, Crogan M, et al. Natural Language Watermarking: Design, Analysis, and a Proof-of-concept Implementation[C]. Proc. of Int. Workshop on Information Hiding. Berlin:Springer-Verlag, 2001: 185-199.

8. YUAN, S. and X. SUN (2006). Design and Implementation of a Multiple Watermarking Algorithm for English Texts [J]. Computer Engineering 15: 051.